Chapter 1

1. Question 6: What are the differences between statistical and practical significance? Is one a prerequisite for the other?

- Statistical significance means that the results are unlikely due to chance alone. Practical significance means that the results are useful or interesting from a real world (economic, scientific, personal) perspective. Statistical significance is prerequisite for practical significance, because without statistical significance it’s unclear that any observed effect is not due to merely chance.

2. Question 7: What are the implications of low statistical power? How can power be improved if it is deemed too low?

- The implication of low power is that the researcher may fail to find significance when it actually exists. This has implications for reproducibility, even in experiments where significance is found. If investigators reject the null hypothesis in an experiment with low power, there’s a good chance that other researchers attempting to reproduce the results will fail to reject, due to low power. Power can be improved by increasing the sample size or controlling for more variation in the experimental design.

3. Question 8: Detail the model-building approach to multivariate analysis, focusing on the major issues at each step.

- **Stage One: Define the research problem, objectives, and multivariate technique to be used.**
  The starting point for any analysis is to define the research problem and objectives in conceptual terms before specifying any variables or measures. This will lead to an understanding of the appropriate type of techniques, dependence or interdependence, needed to achieve the desired objectives. Then based on the nature of the variables involved a specific technique may be chosen.

- **Stage Two: Develop the analysis plan.**
  A plan must be developed that addresses the particular needs of the chosen multivariate technique. These issues include sample size, type of variables (metric vs. nonmetric), and special characteristics of the technique.

- **Stage Three: Evaluate the assumptions underlying the multivariate technique.**
  If the technique requires assumptions of normality, homogeneity of variance, and linearity these assumptions must be checked and validated.

- **Stage Four: Estimate the multivariate model and assess overall model fit.**
  Once the assumptions are validated the model is fit to the data and the goodness-of-fit evaluated. Any corrections that need to be made are made and the fit reassessed.

- **Stage Five: Interpret the variate**
  With acceptable model fit, interpretation of the model reveals the nature of the multivariate relationship.

- **Stage Six: Validate the multivariate model**
  The attempts to validate the model are directed toward demonstrating the generalizability of the results, usually with cross validation or a “test” data set.

Chapter 2

1. Question 3: Distinguish between data which are missing at random (MAR) and missing completely at random (MCAR). Explain how each type will impact the analysis of missing data.

   (a) **Missing at random:** If the missing values of Y depend on X, but not on Y, the data are missing at random. This occurs when X biases the randomness of the observed Y values, such that the observed Y values do not represent a true random sample of actual Y values in the population.

   (b) **Missing completely at random:** When the observed values of Y do not depend on X or on Y and therefore are a truly random sample of all Y values.
(c) Impact: When the missing data are missing at random, the analyst should use only model-based approaches that account for the underlying processes of the missing data for imputation. When the missing data are missing completely at random, the analyst may use simpler methods for imputation. However, multiple imputation methods account for the increased variability and are best used when available for any type of missing data.

2. Question 4: Describe the conditions under which a researcher would delete a case with missing data versus the conditions under which a researcher would use an imputation method.

- The researcher must first evaluate the randomness of the missing data process. If the data are missing at random, deletion or model-based imputation methods must be used. Otherwise, any simple imputation method will reflect the missing data process, resulting in bias.

If the missing data are MCAR, then the decision should be based on theoretical and empirical considerations. If the sample size is large, deletion may be the easiest and most conservative choice, especially for cases with a high number of missing variables. Cases with missing dependent variables are often deleted. If the sample size is small, the analyst may wish to use an imputation method.

3. Question 5: Evaluate the following statement: In order to run most multivariate analyses, it is not necessary to meet all assumptions of normality, linearity, homoscedasticity, and independence.

- This statement is true to a more or less degree depending on the method. Techniques vary in their assumptions and their robustness to assumptions. For example, multiple regression is sensitive to violations of all four assumptions, while discriminant analysis is sensitive only to the assumption of normality, and then only when attempting to draw inferences.
Introduction

Life expectancy is tragically disparate across the nations of the world, with the highest life expectancies being almost twice that of the lowest. For example, Japan has the highest life expectancies: 75.9 years for males and 81.8 years for females, while Malawi has the lowest life expectancies: 38.1 for males and 41.2 for females.

This analysis is concerned with predicting life expectancy by gross national product and birth rate. The data consist of reported birth rates, death rates, infant death rates, life expectancies for males and females, and Gross National Product for 97 countries. The data are from the U.N.E.S.C.O. 1990 Demographic Year Book. We are interested in predicting average life expectancy from gross national product (GNP) and birth rate.

Methods

Average life expectancy was calculated as an average of the reported male and female life expectancies for each country. Exploratory analyses including matrix scatterplots, 3-D scatterplots, and histograms of the residuals were performed to assess the assumptions of the linear regression model. Corrective transformations of birth rate and GNP were taken to linearize the relationship with life expectancy.

Missing data were imputed using all variables in the data set that had a strong and approximately linear relationship with the missing variable, which was exclusively GNP. Assessment of the underlying missing data process was performed with simple inspection for patterns and extent/degree.

All statistical analyses were all performed in SAS.

Results

Exploratory Analyses
life expectancy and both GNP and birth rate. We corrected this using the square transformation for birth rate (birth2) and the natural log transformation for GNP (logGNP).

A three dimensional scatterplot of the transformed predictor variables against average life expectancy showed a fairly nice linear relationship as well.

Residual Analysis

A regression preliminary to the multiple imputation was performed to test the normality of the residuals. A histogram of the residuals was nicely bell-shaped and the various normality tests performed in SAS using Proc Univariate were all non-significant (p-values from 0.13-0.15), indicating that the normal assumption is met for this analysis.

Multiple Imputation

To ascertain which variables to include in the multiple imputation, a scatterplot of all the metric variables was produced using Proc Insight and is shown. It was determined that the relationship between the square of birth rate, death rate, infant death rate and the log of GNP were all suitable for inclusion in the multiple imputation. The SAS output of the multiple imputation of the log of GNP showed that the fraction of missing information was 0.002755, the relative efficiency was 0.99945, and the relative increase in variance for the log of GNP was 0.002759. Proc MIanalyze was used on the complete (imputed) data to estimate the regression parameters and their variances. The results are shown below.

Multiple Imputation Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Error</th>
<th>95% Confidence Limits</th>
<th>DF</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>54.534056</td>
<td>3.129907</td>
<td>48.3990 60.66911</td>
<td>13493</td>
<td>54.006710</td>
<td>55.044832</td>
</tr>
<tr>
<td>birth2</td>
<td>-0.007880</td>
<td>0.000682</td>
<td>-0.00922 -0.00654</td>
<td>35203</td>
<td>-0.007963</td>
<td>-0.007809</td>
</tr>
<tr>
<td>logGNP</td>
<td>2.314285</td>
<td>0.340078</td>
<td>1.64767 2.98090</td>
<td>10568</td>
<td>2.255225</td>
<td>2.373160</td>
</tr>
</tbody>
</table>

All parameters in the model were significant, with p-values less than 0.0001


Conclusions

Both the square of birth rate and the natural log of GNP were highly predictive of life expectancy. Life expectancy will decrease on average by 11.55 for every unit increase in the square of the birth rate. The average life expectancy will increase by 6.81 for every unit increase in the natural log of the gross national product. It appears that decreasing the birth rate and increasing the gross national product could have a beneficial effect on average life expectancies.

Bibliography


Scoring

Chapter Problems

The chapter problems portion of the homework is worth 60 points, or 10 points each. Grading of this section is unavoidably somewhat subjective. Your score is based on how close you came to stating the basic ideas.

Data Analysis

The data analysis is worth 40 points. You started out with the full 40. Points were deducted as follows:

-6 if missing the Introduction, Methods, Results, Conclusions sections. No points were taken off if some of these sections were combined. They just all had to be there somewhere.

-4 if missing scatterplots.

-2 if missing histogram of residuals from preliminary regression model fit.

-3 if missing some statement about the extent and type of missing data process.

-1 if not stated that the data are missing at random.

-2 not identifying that GNP needs a log transformation.

-2 not identifying that birth rate needs some kind of transformation.

-1 not choosing the square transformation.

-5 if no multiple imputation performed.

-2 if no scatterplots to assess which variables to include in the multiple imputation.

-1 if birth (transformed), death, infant death, and LE were not included as variables in the multiple imputation of log GNP.

-1 for not stating which variables were used.

-5 if a confidence interval or a standard error from the MI were not reported with the parameter estimate.

You got 5 points just for turning something in.